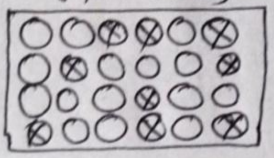
Population and Sample

→ Example – Average height of all the people in the state.
population → state. Average → Mean Population mean $= (\mu) = \frac{1}{1M}\sum_{i=1}^{1M} x_i$

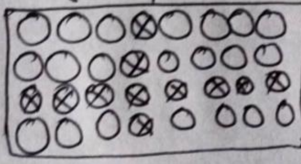
Suppose population of the state is 1million. It is not possible to calculate all 1M.
So we use sample, which is a part of population.
Suppose, sample we take 10,000/10K. Sample mean $= (\bar{x}) = \frac{1}{10K}\sum_{i=1}^{10K} x_i$.

→ Sampling techniques which are grouped in 2 categories – I) Probability Sampling.
II) Non-probability sampling.

Probability Sampling (Randomized)



Non probability Sampling (Non-Random)



Difference of sample selection is based on randomized or not. With randomization, every element gets equal chance to be picked up and to be part of sample for study.

Probability Sampling → Uses randomization to make sure every element of population get equal chances to be part of sample. Also known as Random Sampling.

1) Simple Random sampling – Every element has an equal chance of getting selected to be part of sample. It is used when we don't have any kind of prior information about target population
For eg → Random select of 20 students from class 50 students. P(each student) = 1/50.

2) Stratified sampling – This technique divides the elements of population into small subgroup or strata based on similarity. We need to have prior information about population to create subgroups. For eg → ~~Cluster~~ Strata can be identified such as age, sex, location etc.

3) Cluster sampling – Entire population is divided into clusters or sections then cluster are randomly selected.
a) Single stage clustering sampling → Each cluster is randomly selected for clustering.
b) Two stage clustering sampling → First we randomly select clusters and then from those selected clusters we randomly select elements for sampling.

4) Systamatic Clustering – Here the selection of elements is systematic and not random except the first element. Elements of a sample are choosen at regular interval of population. All the elements are put together in a sequence first, then selected. ① 2 3 ④ 5 6 ⑦ 8 9 ⑩ 11 12 ⑬ 14 15 ⑯ 17

5) Multi-stage Sampling – Combination of one or more methods describe above.
For eg – Country can be divided into states, cities, urban, rural and based on similarity merge together to form strata.

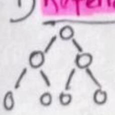
Non-probability Sampling → Does not rely on randomization. Outcomes may be bias.

1) Convenience Sampling – Samples are select based on the availability. It is rare and costly.

2) Purposive Sampling – Only those elements will be selected from population which suits best for purpose of study. For eg – How good tution class is? Show only topper result.

3) Quota Sampling – Elements are selected until exact proportions of certain types of data is obtained or sufficient data in different categories is collected. For example If our population has 45% female & 55% male then our sample should reflect same percentage of female and male respectively.

4) Referral / Snowballing Sampling – This happen when population is completely unknown and rare. Therefore we take help from first element which we select for population and ask him to recommend/say other elements who fit the description. For example, Corona virus. First victim will be catch then the victim will say all possible candidates,



RESER~~VOOI~~R SAMPLING.

Reservoir

Example → Suppose I am in a market for a well balanced cap (hat) to save me from • sun heat.

- So as a buyer, I have the option of selecting one special cap from a large inventory of caps (lot of caps). So tell me, what type of variables will influence my decision?

→ Well as a modern generation, I believe in equal opportunity for all the caps (in the inventory)

→ So lets take a random cap, my first cap. Cap #1 is worthy of my head?

- Hat number = 1, Compatibility probability = $\frac{1}{1}$ = 100% wear the hat Because first and only hat.
- Hat number = 2, Compatibility probability = $\frac{1}{2}$ = 50% [50% keep the hat #1 / 50% wear new hat]

  In this instance, though the second hat had a 50% chance being choosen, it was not.
- Hat number = 3, Compatibility probability = $1/3$ = 33% [wear new hat = 33.3% / Keep current hat = 67.7%]
- Hat number = 4, Compatibility probability = $1/4$ = 25%

  It may seem that the probability of matching is decreasing, but each hat has an equal chance to win.

  So it will go on and on and the probability will decrease but each hat has an equal opurtunity to get selected.

→ Well Reservoir sampling technique is applicable when the sample size is unknown and known as well.

So consider the (i)th hat, with its compatibility probability of $1/i$.

The probability I will be wearing hat i at the time $n > i$ can be demonstrated by a simple formula.

$$\frac{1}{i} * \left(1-\frac{1}{i+1}\right) * \left(1-\frac{1}{i+2}\right) * \ldots\ldots * \left(1-\frac{1}{n}\right) \text{ — } (i)$$

↑ Probability of (i)th hat will be accepted

↑ Probability the $(i+1)$th hat will not be accepted because $(1 - \text{probability})$

↑ Probability the $(i+2)$th hat will not be accepted

↑ Probability the n(th) hat will not be accepted.

Now if we ~~simply~~ simplify equation (i)

Equation (1) is ~~$\frac{1}{i} * \left(1-\frac{1}{i+1}\right) = \frac{1}{i}\left(\frac{i+1-1}{i+1}\right) = \frac{1}{i}\left(\frac{i}{i+1}\right)$~~

$$= \frac{1}{i} * \left(1-\frac{1}{i+1}\right) * \left(1-\frac{1}{i+2}\right) \cdots \left(1-\frac{1}{n}\right)$$

$$= \frac{1}{i} * \left(\frac{i+1}{i+1}-\frac{1}{i+1}\right) * \left(\frac{i+2}{i+2}-\frac{1}{i+2}\right) \cdots \left(\frac{n}{n}-\frac{1}{n}\right)$$

$$= \frac{1}{i} * \left(\frac{i}{i+1}\right) * \left(\frac{i+1}{i+2}\right) \cdots \left(\frac{n-1}{n}\right)$$

Cancel out each other

$$= \frac{1}{\cancel{i}} * \left(\frac{i}{i+1}\right) * \left(\frac{i+1}{i+2}\right) * \ldots * \left(\frac{n-1}{n}\right)$$

$= \frac{1}{n}$. So in this way, the reservoir sampling algo can be used for choosing a sample from a stream of n items, where n is unknown.

So, now the theory parts..

What is Reservoir Sampling?

→ Reservoir sampling is randomized algorithim that is used to select K out of n samples where n is very large or unknown. For example, reservoir sampling can be used to obtain a sample of size K from a population. This algorithim seleck K elements with uniform probability.

Algorithim →

```
int n=8;
int K=4;
# The array to be sample
int input[] = {1,7,4,8,2,6,5,9}; #input array, 8 elements
int output[] = new int[K]; # output array, blank 4 elements.
int i; #initialize one variable i, no value
# Initializing the output array to first K elements of the input array.
for (i=0 ; i< K ; i++){
    output [i] = input[i]; #→ Output [i] = 1,7,4,8. & i=4
}
int j; #initialize one variable j
line → Random num = new Random(); #will generate a random number.
(A)
# Iterating from K to n-1
~~for(j=1~~ for (j=i, j<n ;j++){          #i values is 4, so j=4
    # Generating a random number from 0 to j
    int index = num.nextInt(j+1);
    # Replacing an element in the output with an element in
    the input if the randomly generated is less than K.
    if (index < K){
        Output[index] = input[j];
    }
}
print (Input, Output)          # Input = [1,7,4,8,2,6,5,9]
                                 Output - [6,7,2,5] this value will
                                 change because of Random()
                                 line (A) but size limit is 4.
```

① Copy the first K elements from input array to output array.

② Iterate from K to n-1 (both inclusive. In each iteration j:

2.1) Generate a random number num from 0 to j.

2.2) If num is less than K, replace the element at index num in the output array with the item at index j in the input array.

STEPS IN RESERVOIR SAMPLING →